WHEN DOES A BENCHMARK SCORE LICENSE A CAPABILITY CLAIM? AN EXECUTABLE THREE-STATE IDENTIFICATION CERTIFICATE

Anonymous authors
Paper under double-blind review

ABSTRACT

A benchmark score is routinely read as evidence that a model has a capability, yet the same score can be produced by a non-capability side channel: acquisition metadata, option form, question templates. We ask when a score *licenses* its capability claim, and turn the question into an *executable* procedure. Given a claim, a frozen model score S, and a pre-declared nuisance class $\mathcal{T}$, our certificate returns FAIL when a witness inflates the score without adding intended evidence, SURVIVE when no admissible witness moves it beyond a pre-registered equivalence margin, and INCONCLUSIVE when the data cannot separate the two at that margin, reported together with the inflation the data *can* resolve. A SURVIVE is earned by a two-one-sided-test bound calibrated against injected channels of known strength, so "no witness found" can never be mistaken for "benchmark is valid". A companion identity delimits scoring-side repair: the maximum effective sample size of any channel-neutralizing reweighting equals a mass-weighted overlap efficiency, so localized channels are cheap to repair, while the corpus-wide near-deterministic channels our failures exhibit can only be disclosed. Applied to seven audit units, six of them the benchmarks' own published headline claims, the verdict changes conclusions in both directions: two claims fail, two are positively preserved, and three are returned INCONCLUSIVE and then separated by a power analysis rather than left open. The certificate is a measurement instrument, not a benchmark-condemning machine.

1 INTRODUCTION

Evaluations of vision–language and video models are inferences, not measurements: a score is read as evidence for a latent capability θ, and that reading rests on assumptions that are rarely stated (Jo & Wilson, 2026). When the assumptions fail, the score still exists but no longer licenses the claim. A growing literature asks what claim a benchmark–protocol pair can support, and has converged on what it lacks: Alaa et al. (2025) call for "a new science of benchmarking, focused on creating principled *tools* to evaluate the construct validity of its benchmark datasets," and Jo & Wilson (2026) for "viewing benchmarks not as fixed datasets, but as experimental protocols whose degrees of freedom must be explicitly modeled." What exists is largely *normative* — validity frameworks (Salaudeen et al., 2025; Freiesleben & Zezulka, 2025) and, across 445 benchmarks, an *operational checklist* reporting that "statistical testing was rarely performed" (Bean et al., 2025). Shortcut audits (Buch et al., 2022; Goyal et al., 2017) do run code, but diagnose one benchmark at a time and return no verdict on a stated claim.

We take the next step and make adjudication *executable and statistical*. Our object is a capability claim—"model M's score S on B demonstrates θ"—together with a *pre-declared nuisance class* $\mathcal{T}$, the set of side channels that, by the claim's own logic, must not change what is being measured. The certificate asks whether a witness drawn from $\mathcal{T}$ can move the score without adding intended evidence, and returns a three-state verdict: FAIL, SURVIVE, or INCONCLUSIVE. Crucially, the SURVIVE verdict is not the absence of a found attack— that would reward weak search—but a pre-registered equivalence statement backed by a calibrated instrument.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

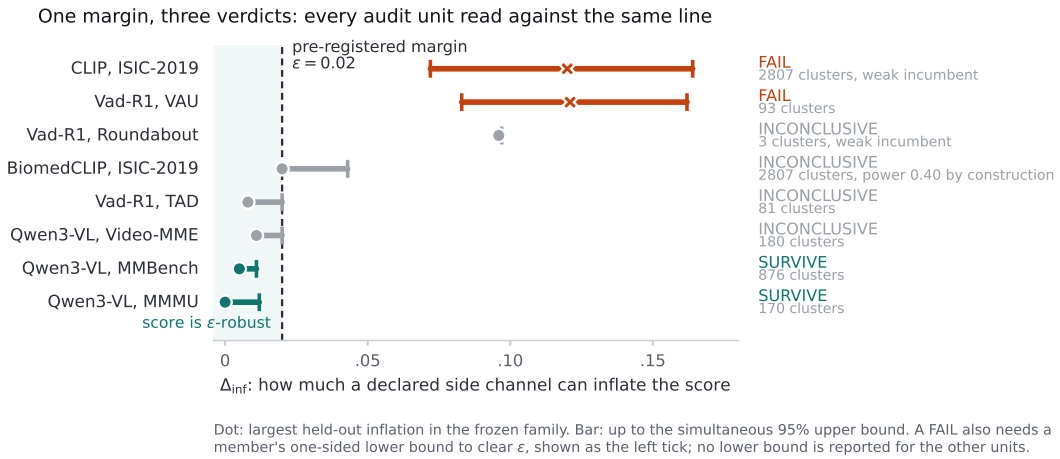


Figure 1: The certificate at a glance. Each benchmark-audit unit's best admissible inflation Δ_{inf} is read against a single margin $\varepsilon = 0.02$, fixed by hash before any verdict is computed. A unit earns SURVIVE only when the simultaneous 95% upper bound on the family maximum falls inside the margin, is returned FAIL only when some member's one-sided lower bound clears it, and is INCONCLUSIVE otherwise. The verdict is therefore a statement about what this data can resolve: Roundabout has three independent clusters and cannot reach either decision by construction, and the two units sitting on the margin are undecided rather than quietly passed. Effective sample size is clusters, not items. The BiomedCLIP row replicates the external unit with a stronger scorer; it is not an eighth unit.

Contributions.

- A formal identification framework (§2): capability claim, declared nuisance class $\mathcal{T}$, and an inflation functional Δ_{inf} estimated by clip-clustered cross-fitting, whose sign and significance define identification.
- A repair–ESS identity (§3): the maximum effective sample size over all channel-neutralizing reweightings equals a mass-weighted per-cell overlap efficiency, yielding a clean dichotomy—localized channels are repairable by down-weighting the offending slice, while corpus-wide near-deterministic channels force $\text{ESS} \to 0$ and can only be disclosed. A constructive demonstration shows a first-moment-balanced repair defeated by a held-out nonlinear attacker that recovers 96% of the inflation on exactly such a global channel.
- A three-state certificate with a calibrated green card (§4–§5): a two-one-sided-test equivalence bound whose false-SURVIVE rate is calibrated against injected channels of known strength.
- An empirical demonstration across seven benchmark-audit units, one fully external and preregistered before scoring (§6): the certificate changes conclusions in *both* directions, and we report its scope honestly—the failures concentrate where exploitable channels exist, while no fresh general-VLM claim is failed — two are licensed and one is left INCONCLUSIVE at this margin.

2 Formal Framework

Claim and nuisance class. A capability claim is a triple (M, B, θ): model M attains score $S = S_M$ on benchmark B, read as evidence for capability θ. The claimant, explicitly or implicitly, commits to a set of transformations that should leave θ unchanged. We make this commitment the primary object.

Definition 1 (Declared nuisance class)**.** *A* nuisance class $\mathcal{T}$ *is a set of item-level side channels Z such that, under the claim, θ does not depend on Z: varying Z while holding the intended evidence fixed must not change what the score is read to mean. A channel is* intended *if θ genuinely depends on it, and* ambiguous *if predictiveness alone cannot decide.*

The taxonomy is normative and claim-relative: the same variable can be nuisance under one claim and intended context under another. For a claim of "understanding the content of a given video," external duration metadata is nuisance; for a claim of "long-video capability," duration is task regime and leaves $\mathcal{T}$. Stating $\mathcal{T}$ before looking at data is what makes the verdict a test rather than a rationalization.

Inflation functional and identification. Let Q_w be the benchmark's native metric under an item weighting w (uniform unless reweighted), g_0 a predictor using only the frozen score S, and g a predictor using S together with a channel $Z \in \mathcal{T}$. Define the *inflation*

$$\Delta_{\inf}(w; M, g) \;=\; Q_w\big(g(S_M, Z)\big) \;-\; Q_w\big(g_0(S_M)\big). \tag{1}$$

Both predictors are fit out-of-fold with *clip-clustered* cross-fitting (items sharing a video or scene never split across folds), and $\Delta_{\inf}$ is always reported on held-out folds. To avoid rewarding an overfit attacker, the admissible attack is the best over a fixed family *including the empty attack* (g_0 itself). The best-admissible attack is selected on training folds and its inflation reported on held-out folds, so a reported $\Delta_{\inf}$ can be slightly negative under sampling noise even though the population maximum is non-negative; we clip at zero only when reporting the point estimate, not the interval. A statistically valid SURVIVE requires that attacker *selection* not contaminate the certification interval; we therefore lock the selected attacker on the design folds (discovery+selection) and compute $\Delta_{\inf}$ with its interval on an independent certification fold (§6), so the reported certificate is the one the theory defines.

Definition 2 (Identification)**.** *θ is* identified from S under $\mathcal{T}$ *if no admissible witness $Z \in \mathcal{T}$ attains $\Delta_{\inf} > 0$ beyond sampling error; a* witness *is any Z with $\Delta_{\inf}$ significantly positive.*

What this estimand is, and what it is not. $\Delta_{\inf}$ is observational and conditional on M: it asks whether Z carries label information that S does not already carry. It measures *shortcut availability above the frozen score*, not whether M's score was produced by the shortcut. A model that has fully absorbed a nuisance channel leaves $\Delta_{\inf} \approx 0$ and is returned SURVIVE, because nothing is left for a witness to add. So SURVIVE reads as "no declared channel can inflate this score further", not "this score is free of the channel". Deciding the latter needs an interventional arm that perturbs Z in M's input and reads the change in S, as CircularEval does for answer position; we do not run one, and §8 returns to it.

3 Disclose, Do Not Repair: A Repair–ESS Identity

A natural temptation, once a witness is found, is to *repair* the benchmark on the scoring side: reweight items by w so that the channel can no longer inflate the score, changing no item and moving w as little as possible. We show exactly when it is and is not achievable. The obstruction is the familiar one from causal inference, where limited overlap bounds what any reweighting can estimate (Crump et al., 2009) and the efficiency of a balancing weight is governed by an overlap quantity (Li et al., 2018). Our contribution is not that fact but its exact evaluation-side form: the maximum effective sample size of a channel-neutralizing reweighting equals a mass-weighted overlap efficiency, which converts a known limit into a computable budget for benchmark repair.

Proposition 1 (Repair floor under a distortion budget)**.** *Reweighting by moment balancing (Hainmueller, 2012) that constrains the first k moments of Z within S-strata neutralizes the linear attackers characterized by those moments, but not higher-order dependence. Rendering Z uninformative about Y given S against an* unrestricted *attacker requires equalizing the full conditional law $P(Z \mid Y, S)$ across Y within strata. Two regimes follow.* (a) Positivity fails (near-deterministic channel). *If some cell (Z, Y, S) carries mass under one label but essentially none under the other, exact neutralization must drive the offending items' weights to zero, so $\mathrm{ESS}(w) \to 0$: the channel can be disclosed but not repaired within any budget—precisely the near-deterministic regime (e.g. a resolution rule that separates classes almost perfectly).* (b) The exact tradeoff is governed by where the channel lives, not only how deterministic it is. *Theorem 1 (Appendix B) gives an* identity *for the maximum effective sample size over all Z-neutralizing reweightings: $\mathrm{ESS}^\star/n = \sum_z f_z\, \psi(\pi_z)$, where f_z is cell z's mass fraction, π_z its label rate, and $\psi(\pi_z) \in [0, 1]$ a per-cell efficiency that is 1 for calibrated cells and vanishes as the cell becomes label-deterministic. Two corollaries delimit repair.* Localized channels are repairable*: if the low-overlap cells carry mass fraction β, the optimal repair simply down-weights that region—the weight-space analogue of item pruning—and retains $\mathrm{ESS}^\star/n \geq (1 - \beta)$, so a*

small poisoned slice never blocks repair. Global near-deterministic channels are not*: when the low-overlap region covers essentially all mass ($\beta \to 1$ and $\eta \to 0$), $\mathrm{ESS}^\star/n \to 0$ and no fixed budget suffices—disclosure, not repair. Unrepairability therefore requires* both *conditions, matching the empirical case: the acquisition channels we fail are corpus-wide rules, not localized artifacts.*

Constructive evidence. We implemented the minimal repair (entropy balancing that decorrelates Z from Y while preserving S–Y) and audited it under a pre-registered protocol. Against the fitted linear attacker the repair drove Δ_{inf} negative, but a held-out nonlinear attacker (gradient-boosted trees / small MLP), never seen by the repair, recovered 96% of the original inflation on the strongest test benchmark (Δ_{inf} from $+0.0884$ raw to $+0.0845$ repaired-held-out, n=439), and on one of the other two video-anomaly benchmarks the held-out attacker ended *above* the unrepaired baseline ($+0.375 \to +0.431$). Meanwhile the reweighting collapsed the effective sample size ($\mathrm{ESS} = 55\%$) and over-corrected a benchmark that carried no channel to begin with (raw $\Delta_{\mathrm{inf}} = +0.005$, driven to -0.007 at $\mathrm{KL} = 0.29 \gg$ the 0.01 budget). A canary channel of known strength ($\gamma = 0.8$) injected into a clean benchmark was detected ($+4.98$ points, $\sim 8\sigma$) and neutralized (-0.65 points held-out), confirming the instrument works and the failure is the method's, not the pipeline's. The lesson is a design principle: a broken benchmark is to be *disclosed by a certificate*, not *repaired at the scoring table*.

The identity is stated for binary labels, and the multi-class extension is open. The natural comparison for the repairable case is adversarial filtering (Le Bras et al., 2020), which prunes rather than reweights; the identity says what budget either has to work with.

4 The Three-State Certificate

Fix a pre-registered practical-equivalence margin ε per benchmark (§5); our demonstration uses a common $\varepsilon = 0.02$ in each benchmark's native metric. Let $[\,\mathrm{LCB}, \mathrm{UCB}\,]$ be a clip-clustered bootstrap confidence interval for Δ_{inf} of the best admissible witness. The certificate returns

$$\begin{cases} \textsc{fail} & \text{if } \exists\, g \in \mathcal{G} :\ \mathrm{LCB}_{95}(\Delta_g) > \varepsilon, \\ \textsc{survive} & \text{if } \mathrm{UCB}_{95}\big(\max_{g\in\mathcal{G}} \Delta_g\big) < \varepsilon \quad \text{(simultaneous, via the bootstrap max statistic)}, \\ \textsc{inconclusive} & \text{otherwise.} \end{cases} \tag{2}$$

The green card is *family-wise*: certifying only the design-selected attacker would leave open that a different member of the frozen family $\mathcal{G}$ exceeds the margin on the population, so SURVIVE requires a simultaneous upper bound on the family maximum, evaluated with every frozen member re-fit on design folds and read on the certification fold. The asymmetry is deliberate. A FAIL is a positive finding—a constructed witness—whose significance we assess by clustered bootstrap bounds; replacing these with an anytime-valid e-process (Zhou et al., 2026) would allow sequential auditing without multiple-testing penalties and is a natural extension we do not implement here. A SURVIVE is an *equivalence* claim and must clear a two-one-sided test, never the mere absence of a found attack. INCONCLUSIVE is reported, not hidden: it says the experiment lacked power, which is information, not a green light.

5 Earning the Green Card: Calibrated Equivalence

SURVIVE is the hard verdict because "I did not see a hole" can mean the hole is small or the eyes are poor. We make it an equivalence test with a controlled error rate, then calibrate what it licenses.

The equivalence test. Fix a margin ε (the largest inflation deemed harmless) before any verdict; here $\varepsilon = 0.02$ in native metric. The declaration files and their hashes are in the supplement. Because that choice is ours, Appendix Table 2 reports every verdict across $\varepsilon \in [0.01, 0.10]$: the two FAILs survive a five-fold widening and no unit ever crosses from FAIL to SURVIVE. Let $\mathrm{UCB}_{1-\alpha}$ be the one-sided clip-clustered bootstrap upper bound on the best-admissible Δ_{inf}.

Proposition 2 (Bounded false-SURVIVE)**.** *Awarding* SURVIVE *iff* $\mathrm{UCB}_{1-\alpha}(\Delta_{\mathrm{inf}}) < \varepsilon$ *is the upper arm of a two-one-sided test of the equivalence null $H_0 : \Delta_{\mathrm{inf}} \geq \varepsilon$. One caveat belongs here rather than in a footnote. The family contains the empty attack, so $\Delta_{\max} \geq 0$ by construction, and on*

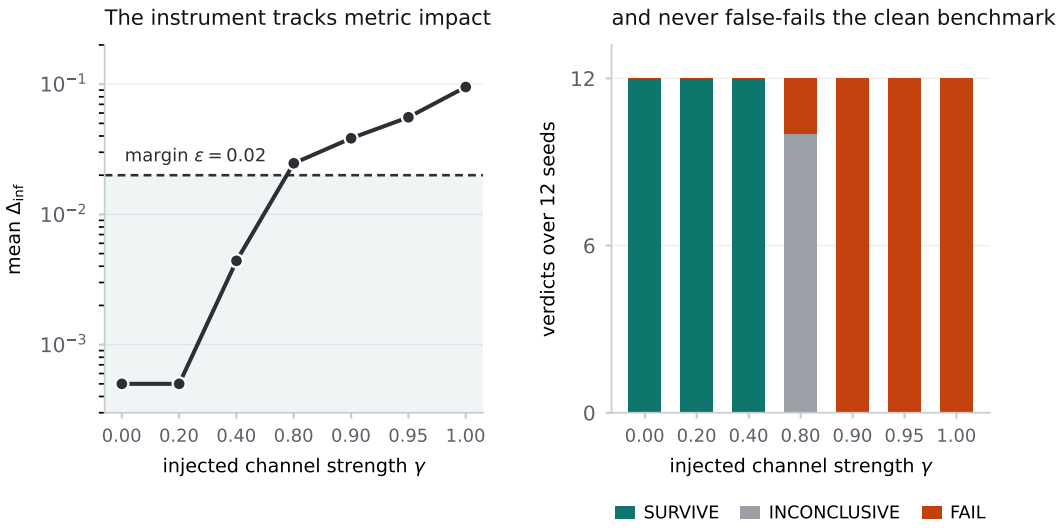


Figure 2: Canary calibration at $\varepsilon = 0.02$, twelve seeds per strength. A channel of known strength γ is injected into MMBench and the certificate is run against it. Left: the mean inflation crosses the margin between $\gamma = 0.4$ and $\gamma = 0.8$. Right: the verdicts follow it, with no FAIL on the clean benchmark and no SURVIVE once the margin is crossed. The left panel also shows what SURVIVE does and does not certify: detection is tied to *metric impact*, so a channel that leaks the answer on a sizeable share of items but moves the metric by 0.0044, as at $\gamma = 0.4$, is certified ε-harmless. That is the intended reading of the margin, and why §5 states the verdict as a property of what the incumbent still gets wrong. The strength axis is categorical.

the units that survive the maximum sits at or near that bound. A bootstrap for a parameter on the boundary of its space is not in general consistent (Andrews, 2000), which is the regime our SURVIVE *verdicts occupy. We did not compute a Bonferroni bound over the K family members, which would avoid the issue, and we report the verdicts as resting on the bootstrap stated below rather than claiming the guarantee is unaffected. Under a consistent bootstrap this rule rejects H_0—and thus certifies* SURVIVE*—with probability at most α whenever the true inflation is $\Delta_{\text{inf}} \geq \varepsilon$. Hence the rate of certifying a benchmark whose channel genuinely moves the metric by at least ε is controlled at α;* SURVIVE *certifies ε-robustness of the score, not the absence of a channel.*

When UCB crosses ε but the point estimate is small the verdict is INCONCLUSIVE: the tape is not sharp enough to rule.

Calibration against known holes. Proposition 2 bounds error only for channels that move the metric past ε; the operational question is which injected channel strengths a strong-model benchmark's metric actually exposes. We inject a channel of known strength γ (leaking the answer with probability γ) into a clean benchmark (MMBench, n=4377) and read the verdict over 12 injection seeds each with a 400-replicate clip-clustered bootstrap over seven strengths (Figure 2).

Two properties matter, and a third qualifies them. The instrument never false-FAILs the clean benchmark, so a SURVIVE is not a trigger-happy pass; its detection power is tied to metric impact; and that power scales with the incumbent's *residual error rate*, not with the channel's strength. At $\gamma = 0.4$ the injected channel leaks the answer on a substantial share of items yet moves the metric by 0.0044, because a strong scorer already answers most of them. What ε-robustness certifies is a property of what M still gets wrong, so a reader must interpret SURVIVE as "robust to any channel that would move the metric by ε," with ε chosen to make that the operative harmlessness threshold. The certificate is computed under the discover–lock–certify protocol of §6: the attacker is selected on design folds, locked, and evaluated on an independent certification fold, so the SURVIVE equivalence bound is not contaminated by attacker selection. A prerequisite throughout is power: before freezing we compute the best-case pass probability of each verdict under a perfectly-working method given

Table 1: Certificate verdicts at the *benchmark-audit-unit* level ($\varepsilon = 0.02$, native metric). †This INCONCLUSIVE reflects both a margin-sized effect and low power ($\Pr(\text{SURVIVE} \mid \Delta_{\text{inf}}{=}0) = 0.40$): maximising an AUC difference over $K{=}6$ family members gives an upper reach of 1.17ε, so this unit could not have delivered SURVIVE on this data. Effective sample size is the number of *clusters*, not items: identical to n_{cert} for VAU, TAD, MMBench and MMMU, but 180 for Video-MME, 2807 for ISIC, and only 3 for Roundabout (see Appendix Table 3). Each unit is one (model, benchmark) pair with its declared nuisance class; Δ_{inf} is the best-admissible inflation with a clip-clustered bootstrap interval. The rightmost column lists the capability claims each unit adjudicates (a claim's verdict is its unit's verdict). Verdicts are *family-wise* (Eq. 2): all K frozen family members are fit on the design folds and read on the independent certification fold (n_{cert}); $\Delta_{\max}$ is the largest held-out inflation in the family and $U_{\max}$ the simultaneous 95% upper bound on the family maximum. A FAIL additionally requires some member's one-sided lower bound to clear ε (VAU $+.083$; ISIC $+.072$). The final row is an *external* unit: a domain (dermoscopic diagnosis) untouched by any of our prior analyses, selected mechanistically for its documented acquisition structure and pre-registered (claim, nuisance class, margin) before any model score was computed. Subgroup-specific claims (e.g. a rule-violation slice) require a subgroup-level estimand and are deferred rather than assigned the unit value. ^{w} marks weak-incumbent units (frozen scorer near chance), where Δ_{inf} conflates channel strength with scorer weakness. The BiomedCLIP row is a pre-registered replication of the external unit with a stronger scorer, not an eighth unit; the seven units are the rows above it.

Unit (model, benchmark)	n_{cert}	K	$\Delta_{\max}$ / $U_{\max}$	Verdict	Claims adjudicated
Vad-R1, VAU	93	7	$+.121$ / .162	FAIL	anomaly reasoning; understanding
Vad-R1, Roundabout^{w}	74	5	$+.096$ / .097	INCONCLUSIVE	roundabout detection
Vad-R1, TAD	81	7	$+.008$ / .020	INCONCLUSIVE	discrimination ranking
Qwen3-VL, MMBench	876	3	$+.005$ / .011	SURVIVE	perception & reasoning; ranking
Qwen3-VL, MMMU	170	3	$+.000$ / .012	SURVIVE	college / multi-discipline reasoning
Qwen3-VL, Video-MME	540	3	$+.011$ / .020	INCONCLUSIVE	video understanding; ranking
CLIP, ISIC-2019^{w}	5066	6	$+.120$ / .164	FAIL	skin-lesion diagnosis (external)
BiomedCLIP, ISIC-2019	5066	6	$+.020$ / .043	INCONCLUSIVE†	skin-lesion diagnosis (external)

the interval width, and a benchmark whose interval is wider than the effect it must resolve is reported INCONCLUSIVE rather than allowed to produce a spurious verdict.

6 Empirical Demonstration

We apply the certificate to seven benchmark-audit units spanning *four* distinct scorers (Vad-R1, Qwen3-VL, CLIP, BiomedCLIP) — three units share the Vad-R1 checkpoint across three different benchmarks, so the seven units are not seven independent models, and we count them as seven *claim–benchmark pairs* rather than seven replications: three video-anomaly units (VAU, TAD, Roundabout; scorer Vad-R1-7B), three general vision–language units (MMBench, MMMU, Video-MME; scorer Qwen3-VL-8B), and one external medical-imaging unit (ISIC-2019; scorer CLIP ViT-B/32, zero-shot). The claim pool is fixed mechanically—each benchmark's own headline capability statement instantiated on its frozen scorer—and fixed before adjudication; the external unit's claim, nuisance class, margin and splits were declared before any of its scores existed. The declarations ship with the supplement, with the hashes that show they are unchanged since they were written; we claim no external timestamp. All verdicts follow the discover–lock–certify protocol: the attacker is selected and locked on design folds and Δ_{inf} is computed with clip-clustered bootstrap intervals on the independent certification fold. Nuisance classes follow each unit's declared claim (acquisition metadata for anomaly and medical units; option form / template / duration for general VLM), with ambiguous channels excluded from FAIL eligibility.

Figure 1 places every unit against the margin and Table 1 reports the seven under the actual discover–lock–certify protocol on independent certification folds: two FAIL (VAU; ISIC-2019), two SURVIVE (MMBench, MMMU), three INCONCLUSIVE (Roundabout, TAD, Video-MME). We report at the unit level deliberately—the adjudicated claims collapse onto seven independent measurements, so

counting claims would multiply-count a single witness. Four observations. First, the certificate *preserves* as well as fails: every fresh general-VLM claim clears the equivalence bar or is honestly marked underpowered, so the instrument is not a benchmark-condemning machine. MMBench's green card is comfortably powered ($\Pr(\text{SURVIVE} \mid \Delta_{\text{inf}}{=}0) = 0.98$); MMMU's sits on the conventional threshold (0.78) and is licensed at $\varepsilon = 0.02$ but would be underpowered at any tighter margin, its 170 items admitting no inflation smaller than 0.0059. Second, INCONCLUSIVE does its job on the independent fold: TAD and Video-MME both land with $U_{\max} = .020$, exactly on the margin, so neither is licensed at $\varepsilon = 0.02$ (both cross to SURVIVE only at $\varepsilon = 0.03$, Appendix Table 2); their small positives are left unresolved, and Roundabout's point estimate is large ($+0.096$) but its few scene clusters leave the lower bound at zero, so the certificate declines to fail rather than over-claiming on a wide interval. The power analysis (Appendix Table 3) sharpens the second case into a boundary result we prefer to state than to omit: Roundabout's 74 items form only *three* scene clusters, and with 73 positives to one negative the estimand's finest representable step is $0.0137 = 0.69\varepsilon$. Under the certificate's own bootstrap its verdict is then near-independent of the truth — across true inflations from 0.014 to 0.096 the three verdicts arrive at roughly equal rates, the reported INCONCLUSIVE reproduces only 36% of the time, and this is the single arm where the calibration guarantee does not hold empirically ($\Pr(\text{FAIL}) = 0.33$ at a truth below ε). We keep the unit as a worked example of the instrument's boundary: what makes a verdict trustworthy is the number of independent clusters, not the item count, and the diagnostic that detects the difference is part of the procedure rather than an afterthought. Third, the external unit is a *transportability witness*, and its role must be stated precisely. That skin-lesion benchmarks harbor acquisition artifacts is established dermatology-AI literature: models classify ISIC lesions with the lesion pixels removed (Bissoto et al., 2019; 2020), and surgical ink markings alone shift CNN melanoma calls (Winkler et al., 2019). Our contribution on this unit is therefore not the channel's existence but three things the prior literature does not do: (a) a *claim-level, pre-registered adjudication*—claim, nuisance class, margin, and splits frozen before any score existed—rather than a dataset-level post-hoc demonstration; (b) a *score-conditioned estimand*: prior work shows the channel predicts labels ($I(Y;Z) > 0$), whereas the certificate measures what the channel adds *above the frozen model's score*, which is what a leaderboard reading needs; and (c) a validation of mechanism-driven selection: we chose ISIC by the pre-registered rule "multi-source acquisition structure," and the dermatology literature independently documents the same pathology—convergent evidence that the selection rule tracks a real mechanism. The verdict itself is clean ($\Delta_{\text{inf}} = +0.120$, $\text{LCB} = 0.064 \gg \varepsilon$): four zero-pixel file properties add $+12$ AUC points above the frozen score, and the nuisance-only predictor (0.636) outperforms the model itself (0.523). The unit's claim is *instantiated* by our pre-registered rule (the frozen model's score read as skin-lesion diagnostic capability) rather than quoted from a publication; the dermatology-VLM literature in fact reports generic CLIP zero-shot as weak on dermoscopy (Yan et al., 2025), independently corroborating the weak-incumbent flag. The nuisance declaration is claim-relative, not a claim that these attributes carry no medically relevant information: under $\theta =$ "visual recognition of the lesion content presented in the image," external encoding and container attributes are declared not to constitute intended evidence, whatever else they may correlate with. Fourth, a pre-registered replication of the same unit with a stronger published medical scorer (BiomedCLIP, base AUC 0.686 vs. CLIP's 0.525) shows the channel *collapse*: the family maximum falls six-fold to $+0.020$ and the certificate declines to rule ($U_{\max} = .043$, $\max\text{LCB} = -.005$) rather than over-claiming in either direction. The pair $(\text{CLIP}, B, \mathcal{T}) = \text{FAIL}$ versus $(\text{BiomedCLIP}, B, \mathcal{T}) = \text{INCONCLUSIVE}$ is direct evidence that verdicts bind to the $(M, B, \mathcal{T})$ triple—a stronger incumbent absorbs the channel, exactly as the score-conditioned estimand predicts—and that the external unit's FAIL is a statement about a weak scorer's reading of the benchmark, not a permanent property of the benchmark. A single external unit remains evidence of transport, not of prevalence. On fresh general-VLM units no witness survives certification—the same instrument that certifies several general-VLM benchmarks as resistant prospectively identifies a strong nonvisual witness in an independently selected medical-imaging benchmark.

Why the green cards are green. The surviving units are not lucky: their resistance is mechanistic, which is what a SURVIVE should ultimately trace to. MMBench's CircularEval rotates option positions by design, ablating exactly the answer-position channel our nuisance class declares—a deployed defense, and the certificate confirms it holds ($\Delta_{\text{inf}} = +0.005$). On the video units, the evaluation pipeline's own preprocessing (uniform frame sampling to a fixed budget, resolution normalization) physically severs file-level acquisition metadata from the model's input, so the channel cannot re-enter through the score. And on MMMU a strong model's score already absorbs whatever legitimate signal

option text carries, so the admissible surface form adds nothing ($\Delta_{\text{inf}} = 0.000$): predictive side information in isolation does not imply residual inflation above a strong score. Survival, in other words, is evidence that a defense is deployed or a channel is severed—information a benchmark designer can act on, and the constructive reason the instrument issues green cards rather than merely failing to find witnesses. The contribution is therefore the instrument—the executable, calibrated, both-directional certificate—rather than a claim that leaderboards are broadly invalid.

How the units were chosen, and the one we removed. A demonstration that selects its own test cases can manufacture any failure rate it likes, so we state the enumeration rule and the full disposition. Units were formed by pairing each benchmark on which we already held a frozen scorer with *that benchmark's own published headline claim*; claims were instantiated mechanically from the benchmark's text and hash-frozen before any adjudication, and the nuisance class follows the claim's own logic rather than our expectations about what would fail. Six of the seven therefore reuse benchmarks whose channel structure was known to us a priori. We read those six as *retrospective* — they show the instrument returns the right verdict where the answer is independently known — and not as an estimate of how often benchmarks in general fail. The seventh unit is external in domain and was pre-registered before any of its scores existed.

One unit was removed after it had been adjudicated, and we disclose it because the removal is otherwise invisible. An eighth unit adjudicated a *rank* claim rather than an accuracy claim—a live competition leaderboard, whose questions, scoring script and held-out set were set by a third party—under the same discover–lock–certify protocol and the same hash-frozen split. It returned FAIL: a nuisance-only rule using no video and no model score reached 1.000 on the certification fold (exact one-sided lower bound 0.829 over clusters; the clip-clustered bootstrap degenerates to $[1, 1]$ on an all-correct result and was replaced by the exact bound, a substitution forced by the estimator's known boundary behaviour rather than by the result). We removed it because the same evidence is reported separately, so keeping it here would present one witness twice. The direction of the removal is worth stating plainly: it deletes a FAIL, so the $2/7$ failure rate we report is lower than the rate we measured, not higher. This is the only unit removed after adjudication, and no unit was removed because of which way its verdict fell.

7 Related Work

Our nearest conceptual neighbours are the normative-validity line: Alaa et al. (2025) argue the field needs principled tools for construct validity, and Bean et al. (2025) survey 445 benchmarks and ship an operational checklist. We advance from taxonomy and checklist to an executable, statistically certified adjudication. Jo & Wilson (2026) argue evaluations should be grounded in a theory of capability and show, under a mean-zero-perturbation assumption, that benchmarks violate identification; their intervention is input rephrasing and their output an ability estimate, whereas we admit a declared adversarial nuisance class and output a three-state certificate—answering their call to model benchmarks as experimental protocols. Post-hoc auditing of realized exploits (Shao et al., 2026) quantifies a mislead gap on agentic traces; we differ in operating on perception benchmarks, pre-declaring the nuisance class, and issuing an equivalence-backed SURVIVE. Validity-centered and construct-validity frameworks (Salaudeen et al., 2025; Freiesleben & Zezulka, 2025) supply the normative vocabulary we make executable. Closest in verdict structure, Reynolds (2026) assign registered-tolerance, bounds-based three-state statuses (supported / defeated / unresolved) to *projectibility* of audit inferences in a deployment setting; we cede the three-state bounds-based verdict form to that line, and differ in the object and the machinery: our verdicts adjudicate capability claims against an adversarially searched, pre-declared nuisance class, with selection isolated by discover–lock–certify, the green card calibrated by injected canaries, and a repair–ESS identity delimiting what a FAIL can be turned into. Verifier frameworks in the LM-judge lineage score candidate solutions with continuous feedback; our object is a capability claim about a benchmark score and our output a three-state certificate over a declared nuisance class. Concurrent audit protocols target single failure modes—audit-constrained prompt variation (Li, 2026) and phrasing-controlled textual-prior isolation (Singla et al., 2026)—but issue empirical comparisons or mitigations rather than an equivalence-backed three-state verdict over a declared nuisance class. Our repair analysis instantiates entropy balancing (Hainmueller, 2012), and the limit it runs into is the overlap limit familiar from causal inference (Crump et al., 2009; Li et al., 2018). What we have not found stated for evaluation-side

repair is the exact identity: that the best achievable effective sample size equals a mass-weighted overlap efficiency, which turns that limit into a budget a benchmark designer can compute. This is a non-exhaustive search of the reweighting and covariate-shift literature.

8 LIMITATIONS

The estimand is observational, and an interventional arm is the obvious next step. $\Delta_{\inf}$ asks what a channel adds above the frozen score, so a SURVIVE cannot rule out a model that already reads the channel itself. The test that would is specific: perturb Z in the model's input and read the change in S. We did not run it, and until someone does a SURVIVE bounds residual inflation rather than clearing the model.

The verdicts are a function of ε, and we chose ε. At the declared $\varepsilon = 0.02$ the seven units split two FAIL, two SURVIVE and three INCONCLUSIVE; at $\varepsilon = 0.01$ every unit returns INCONCLUSIVE (Appendix Table 2). Four of the six retrospective verdicts differ between the initial cross-fitted analysis and the final family-wise column (Appendix Table 4), always toward a weaker claim; three of those four moved at the first tightening, when the independent certification fold was introduced. Both facts are properties of a bounds-based instrument rather than defects of this one: an equivalence margin is a declaration of what harm is, and a tighter declaration buys caution at the cost of resolution. But they mean the correct reading of a SURVIVE is "no admissible witness moves this score by two points on its native metric", never "this benchmark is valid", and a reader who disagrees with $\varepsilon = 0.02$ should read Appendix Table 2 rather than this paper's verdict list.

Six of the seven units reuse benchmarks whose channel structure was known a priori, so their verdicts are retrospective; the seventh (ISIC-2019) is fully external and pre-registered, turning the demonstration partially external, though a single external unit is evidence of transportability, not an external prevalence estimate. The repair–ESS identity concerns *minimal* scoring-side reweighting and does not preclude data-side redesign. The nuisance class is a normative declaration, not a discovered fact; a wrong $\mathcal{T}$ yields a well-formed but mis-aimed verdict. Every verdict is bound to the $(M, B, \mathcal{T})$ triple and a single scorer (Vad-R1-7B or Qwen3-VL-8B); a different scorer can change it. The SURVIVE guarantee assumes attacker selection is isolated from the certification data by the discover–lock–certify split, which our demonstration instantiates directly; the cost is smaller certification folds and hence wider intervals, which is why two units are reported INCONCLUSIVE. Finally, the three states are the ones our instrument can separate, not a claim to enumerate every route by which a benchmark score can overstate a capability.

9 CONCLUSION

A benchmark score licenses a capability claim only when no declared nuisance channel can inflate it further above the model's own score. We made that condition executable, and the instrument the field needs is certification rather than repair.

REFERENCES

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Position: Medical large language model benchmarks should prioritize construct validity. In *International Conference on Machine Learning (Position Track)*, 2025. arXiv:2503.10694.

Donald W. K. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–406, 2000.

Andrew M. Bean et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*, 2025. arXiv:2511.04703.

Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (de)constructing bias on skin lesion datasets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? not so fast. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

Timo Freiesleben and Sebastian Zezulka. The benchmarking epistemology: Construct validity for evaluating machine learning models. *arXiv preprint arXiv:2510.23191*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Nathanael Jo and Ashia Wilson. Position: Ai evaluations should be grounded on a theory of capability. *International Conference on Machine Learning (ICML)*, 2026.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International Conference on Machine Learning (ICML)*, 2020.

Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Hongmin Li. Targeted tests for llm reasoning: An audit-constrained protocol. *arXiv preprint arXiv:2605.11599*, 2026.

Brett Reynolds. When benchmark inferences do not compose: Projectibility in AI evaluation. *arXiv preprint arXiv:2607.26159*, 2026.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Benjamin Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*, 2025.

Jiaqi Shao, Hanck Chen, Wei Zhang, Maxm Pan, and Bing Luo. Do agent benchmarks measure capability? protocol validity in the age of agentic AI. *arXiv preprint arXiv:2607.22368*, 2026.

Pratham Singla, Shivank Garg, Vihan Singh, and Paras Chopra. Do vision–language models see or guess? measuring and reducing textual-prior reliance with a phrasing-controlled benchmark. *arXiv preprint arXiv:2606.10400*, 2026.

Julia K. Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A. Haenssle. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, 155(10):1135–1141, 2019.

Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. *arXiv preprint arXiv:2503.14911*, 2025.

Siyu Zhou, Patrick Vossler, Venkatesh Sivaraman, Yifan Mai, and Jean Feng. Adaptive auditing of AI systems with anytime-valid guarantees. *arXiv preprint arXiv:2605.07002*, 2026.

REPRODUCIBILITY STATEMENT

The supplement ships two executable entry points and a set of frozen records. `canary.py` runs the complete certificate against a channel of known strength injected into MMBench, reading only cached option log-probabilities, and reproduces Figure 2 on CPU. `audit.py` replays the three-state decision over the committed family-wise bounds, checks it cell by cell against Table 1, and confirms that the certification fold shipped for each unit is the one its verdict was read on. Refitting the attacker for the remaining units needs each benchmark's own images and question files, which we cannot redistribute, so that step is not executable here; the readings it produced are included as records. Each split file carries a hash of its own contents, which shows the file is unchanged since it was written and nothing more: we claim no external registry and no third-party timestamp.

USE OF AI ASSISTANCE

An AI assistant (Claude, Anthropic) was used throughout this work under the authors' direction: to implement and debug the experiment harness and the cache construction, to write the scripts that generate every table from the committed run records, to run pre-registered analyses on cached tensors, to check claims in this manuscript against those records, and to draft and edit prose, including this section. In the sessions that produced this text the assistant did not contribute to formulating Theorem 1, the propositions, or their proofs. All experimental designs, pre-registration rules, verdicts and decisions about what the paper claims were made by the authors, who read and are responsible for every sentence and every number. Every verdict was computed by committed scripts and decided by the authors.

A VERDICT SENSITIVITY AND POWER

The two tables below support the Limitations section. Table 2 reports what every unit would have received at margins other than the pre-registered $\varepsilon = 0.02$, and Table 3 reports the inflation each INCONCLUSIVE unit is powered to detect, so that an undecided verdict can be read as a resolution limit rather than an absence of evidence.

Table 2: Sensitivity of every verdict to the equivalence margin. The pre-registered value is $\varepsilon = 0.02$; the columns show what each unit would have received at other margins under the same intervals. Both FAIL verdicts are stable across a five-fold range of ε and only dissolve into INCONCLUSIVE at $\varepsilon = 0.10$, five times the registered margin; no unit ever turns from FAIL to SURVIVE. The green cards are the conservative direction: at $\varepsilon = 0.01$ every unit returns INCONCLUSIVE, so a stricter margin costs us evidence rather than manufacturing it.

Unit	LCB	UCB	.01	**.02**	.03	.05	.10
Vad-R1, VAU	.083	.162	FAIL	FAIL	FAIL	FAIL	INCONCLUSIVE
Vad-R1, TAD	.000	.020	INCONCLUSIVE	INCONCLUSIVE	SURVIVE	SURVIVE	SURVIVE
Vad-R1, Roundabout	.000	.097	INCONCLUSIVE	INCONCLUSIVE	INCONCLUSIVE	INCONCLUSIVE	SURVIVE
Qwen3-VL, MMBench	.000	.011	INCONCLUSIVE	SURVIVE	SURVIVE	SURVIVE	SURVIVE
Qwen3-VL, MMMU	.000	.012	INCONCLUSIVE	SURVIVE	SURVIVE	SURVIVE	SURVIVE
Qwen3-VL, Video-MME	.002	.020	INCONCLUSIVE	INCONCLUSIVE	SURVIVE	SURVIVE	SURVIVE
CLIP, ISIC-2019	.072	.164	FAIL	FAIL	FAIL	FAIL	INCONCLUSIVE

Table 3: Power of the equivalence test, computed under the certificate's own clip-clustered bootstrap ($\varepsilon = 0.02$). Effective sample size is the number of *clusters* n_{clus}, not items. "reach" is $(U_{\max} - \Delta_{\max})/\varepsilon$, how far the one-sided upper bound stretches above the point estimate in units of the margin. $\Pr(\text{SURVIVE} \mid \Delta_{\text{inf}}{=}0)$ is the probability of correctly awarding the green card to a channel-free unit, from a nested bootstrap (outer: resample the certification fold's clusters; inner: the certificate's own B-replicate bootstrap and Eq. 2); τ_{80} is the largest true inflation still certified $\geq 80\%$ of the time; the last column is the realised $\Pr(\text{SURVIVE})$ at the achievable inflation nearest the margin, to be read against $\alpha = 0.05$ up to the Monte-Carlo error below; the ‡ entries are read *below* ε, where awarding SURVIVE is correct, so they are not false-SURVIVE rates. Monte-Carlo SE ≤ 0.011 for the two SURVIVE units, ≤ 0.035 elsewhere. ‡ read at the nearest achievable population inflation, given in parentheses. $\Pr(\text{SURVIVE}\,|\,0)$ abbreviates $\Pr(\text{SURVIVE} \mid \Delta_{\text{inf}}{=}0)$.

Unit	$n_{\text{cert}}/n_{\text{clus}}$	reach	$\Pr(\text{SURVIVE}\,\|\,0)$	τ_{80}	$\Pr(\text{SURVIVE})$ at ε	Verdict	Power reading
Vad-R1, VAU	93/93	2.07	$0.30^{\ddagger}$ (.004)	—	.04	FAIL	FAIL powered (1.00)
Vad-R1, Roundabout	74/**3**	0.06	1.00	$< .014$	$.32^{\ddagger}$ (.014)	INCONCLUSIVE	unresolvable: 3 clusters
Vad-R1, TAD	81/81	0.61	$0.84^{\ddagger}$ (.001)	.002	.07	INCONCLUSIVE	genuine mid-range effect
Qwen3-VL, MMBench	876/876	0.34	**0.98**	.007	.00	SURVIVE	adequately powered
Qwen3-VL, MMMU	170/170	0.59	**0.78**	$< .006$	.01	SURVIVE	valid, marginally powered
Qwen3-VL, Video-MME	540/180	0.47	0.82	.001	.02	INCONCLUSIVE	genuine mid-range effect
CLIP, ISIC-2019	5066/2807	2.19	$0.22^{\ddagger}$ ($-.003$)	—	.08	FAIL	FAIL powered (0.96)
BiomedCLIP, ISIC-2019	5066/2807	1.17	$0.40^{\ddagger}$ ($-.001$)	—	.07	INCONCLUSIVE	margin-sized effect *and* low power

B THE REPAIR–ESS IDENTITY: STATEMENT, PROOF, AND COROLLARIES

We work within a stratum $S = s$ and suppress it; the aggregate statement follows by mixing over strata. The channel Z takes values in a finite set of cells. Cell z has n_z items ($f_z = n_z/n$), within-cell positive rate $\pi_z = \Pr(Y{=}1 \mid Z{=}z)$, and the overall target rate is $\bar{\pi}$. Weights $w_i \geq 0$ satisfy $\sum_i w_i = n$ and $\text{ESS}(w) = n^2/\sum_i w_i^2$. Call w *Z-neutralizing* if $P_w(Y{=}1 \mid Z{=}z) = \bar{\pi}$ for every cell with positive reweighted mass—exactly the condition that a Bayes attacker with access to Z gains nothing over the score alone, i.e. $\Delta_{\text{inf}}(w) = 0$ against an unrestricted attacker.

Theorem 1 (Repair–ESS identity)**.** *Define the per-cell efficiency*

$$\psi(\pi) \;=\; \left[\frac{\bar{\pi}^2}{\pi} + \frac{(1-\bar{\pi})^2}{1-\pi}\right]^{-1} \text{ for } \pi \in (0,1), \qquad \psi(0) = \psi(1) = 0, \tag{3}$$

so that $\psi(\bar{\pi}) = 1$, ψ is maximized at $\pi = \bar{\pi}$, and $\psi(\pi) \to 0$ as the cell becomes label-deterministic. Then

$$\max_{w \text{ Z-neutralizing}} \frac{\mathrm{ESS}(w)}{n} = \sum_z f_z \, \psi(\pi_z). \tag{4}$$

Proof. Step 1 (within a cell). Fix cell z with reweighted mass $\mu_z = \sum_{i \in z} w_i$. Neutralization forces the cell's reweighted label composition to $(\bar{\pi}, 1-\bar{\pi})$: mass $\bar{\pi}\mu_z$ on its $n_z\pi_z$ positive items and $(1-\bar{\pi})\mu_z$ on its $n_z(1-\pi_z)$ negative items (for $\pi_z \in \{0,1\}$ this is possible only with $\mu_z = 0$, consistent with $\psi = 0$). Among all weightings achieving these two block masses, $\sum_{i \in z} w_i^2$ is minimized by uniform weights within each (cell, label) block, giving $\sum_{i \in z} w_i^2 \geq \frac{(\bar{\pi}\mu_z)^2}{n_z\pi_z} + \frac{((1-\bar{\pi})\mu_z)^2}{n_z(1-\pi_z)} = \frac{\mu_z^2}{n_z\,\psi(\pi_z)}$, with equality achieved.

Step 2 (across cells). It remains to minimize $\sum_z \mu_z^2/(n_z\psi_z)$ subject to $\sum_z \mu_z = n$, $\mu_z \geq 0$. By Cauchy–Schwarz the minimizer is $\mu_z \propto n_z\psi_z$, with minimum value $n^2/\sum_z n_z\psi_z$. Hence $\max \mathrm{ESS} = n^2/\min \sum_i w_i^2 = \sum_z n_z\psi_z$, which is (4). Both steps are achieved by explicit weights, so the identity is exact, not a bound. □

Corollary 1 (Global near-deterministic channels: disclose, do not repair)**.** *If cells of minority fraction $m_z = \min(\pi_z, 1-\pi_z) \leq \eta$ carry all but a vanishing share of the mass ($\beta \to 1$), then $\psi(\pi_z) \leq m_z/\min(\bar{\pi}, 1-\bar{\pi})^2$ gives $\mathrm{ESS}^\star/n \leq \eta/\min(\bar{\pi}, 1-\bar{\pi})^2 \to 0$: no fixed budget $\mathrm{ESS} \geq \rho n$ admits a Z-neutralizing repair. (The η-only bound of an earlier draft is precisely this $\beta \to 1$ special case; it is false in general—see Corollary 2.)*

Corollary 2 (Localized channels: weight-space pruning repairs)**.** *If the low-overlap region carries mass fraction β while the remaining cells are calibrated ($\pi_z = \bar{\pi}$), the identity gives $\mathrm{ESS}^\star/n \geq 1-\beta$: the optimal repair down-weights the offending region toward zero—the weight-space analogue of item pruning—and a small poisoned slice never obstructs repair at moderate budgets. Repair is infeasible at budget ρ iff $\sum_z f_z\psi(\pi_z) < \rho$, which requires the deterministic region to dominate the corpus.*

Quantitative match. On the failing video-anomaly unit we evaluate the identity directly: cells fixed in advance as the naturally discrete resolution pairs (w, h) on the design split (28 cells) give a ceiling $\mathrm{ESS}^\star/n = 0.582$ for *any* fully neutralizing reweighting—below the pre-registered repair budget $\mathrm{ESS} \geq 0.8n$, so the theorem predicts the repair gate's failure before it is run. The entropy-balanced repair in fact spent down to $\mathrm{ESS} = 0.549$, near the frontier, while matching only first moments—and was still defeated by a held-out nonlinear attacker: infeasible under budget by the identity, and under-neutralizing even at the frontier. *Empirical match.* The dichotomy of Corollaries 1–2 is the one our constructive experiment instantiates: the header/resolution channels we fail are *corpus-wide* acquisition rules ($\beta \approx 1$), and the entropy-balanced repair indeed collapsed to $\mathrm{ESS} = 55\%$ while a held-out nonlinear attacker recovered the inflation; a channel confined to a small slice would instead be removable by down-weighting that slice, consistent with item-pruning repairs in the literature. This also explains why disclosure-based certification, not scoring-side repair, is the right product exactly in the regime our FAIL verdicts occupy.

C Verdict Migration under Protocol Tightening

An earlier analysis assigned verdicts by clustered cross-fitting alone, without an independent certification fold. Introducing the independent certification fold moved three of the six retrospective verdicts, and a fourth moved when verdicts were made family-wise, so four of six differ between the first column and the last; we report the migration rather than silently adopting the final column, because the direction of every move is itself evidence that the protocol works.

Table 4: Verdicts before (cross-fitting) and after (independent certification fold) protocol tightening. Wider intervals on the smaller certification folds move borderline units toward INCONCLUSIVE, never toward a stronger claim: Roundabout's wide scene-clustered interval loses its FAIL, Video-MME's small positive loses its SURVIVE, and MMMU's interval tightens into SURVIVE. The final column applies the family-wise green card of Eq. 2, which withdraws TAD's; the verdicts reported everywhere else in this paper are that column. Every protocol tightening here moves verdicts toward weaker claims or resolved power, never toward stronger ones.

Unit	cross-fitting	certification fold	family-wise (final)	direction
VAU	FAIL	FAIL	FAIL	unchanged
Roundabout	FAIL	INCONCLUSIVE	INCONCLUSIVE	weaker claim (wide interval honestly surfaced)
TAD	SURVIVE	SURVIVE	INCONCLUSIVE	weaker claim (green card withdrawn over the family)
MMBench	SURVIVE	SURVIVE	SURVIVE	unchanged
MMMU	INCONCLUSIVE	SURVIVE	SURVIVE	resolved (interval cleared ε)
Video-MME	SURVIVE	INCONCLUSIVE	INCONCLUSIVE	weaker claim (small positive unresolved)